Known Classes

Expression Data

Class Discovery

Assess Gene-Class
Correlation:
Neighborhood Analysis

Build Predictor

Test Predictor
by Cross-validation

Test Predictor
on Independent Dataset

Figure 1A

$$c = (1,1,1,1,1,0,0,0,0,0,0)$$

$$gene_1 = (e_1, e_2, e_3, \ldots, e_{12})$$

$$gene_2 = (e_1, e_2, e_3, \ldots, e_{12})$$

Figure 1B

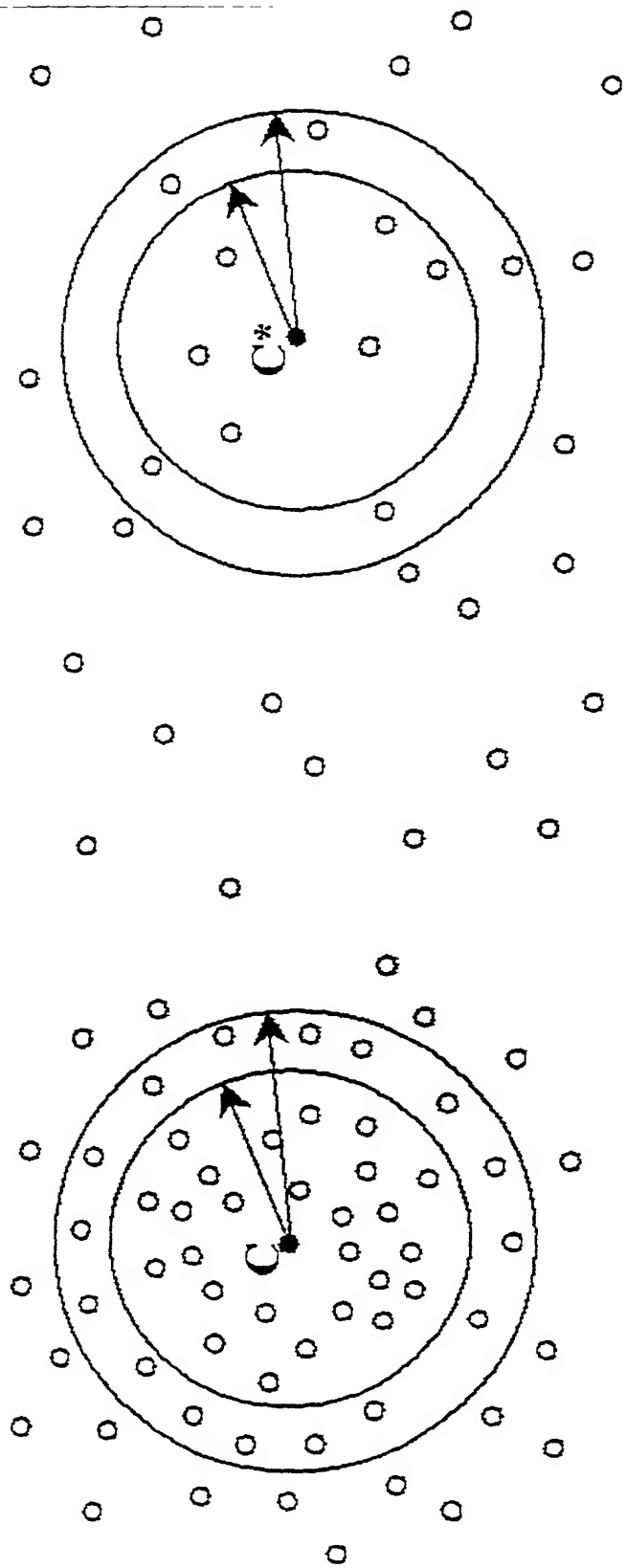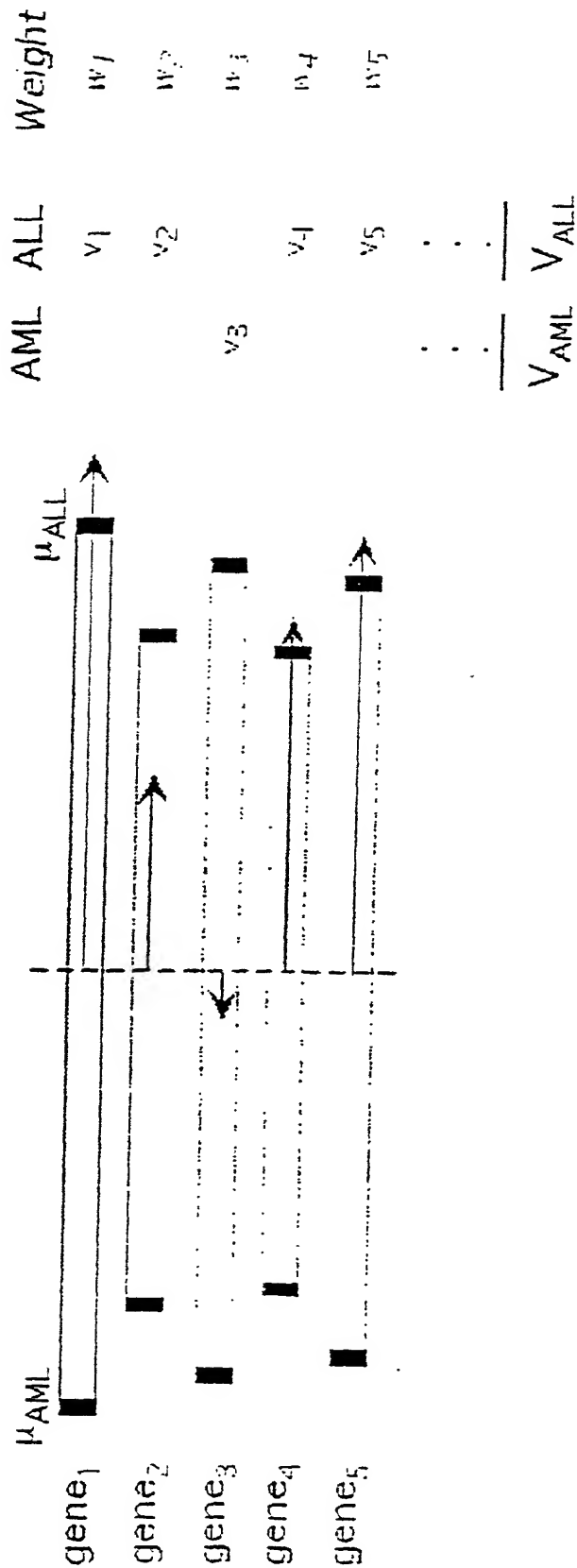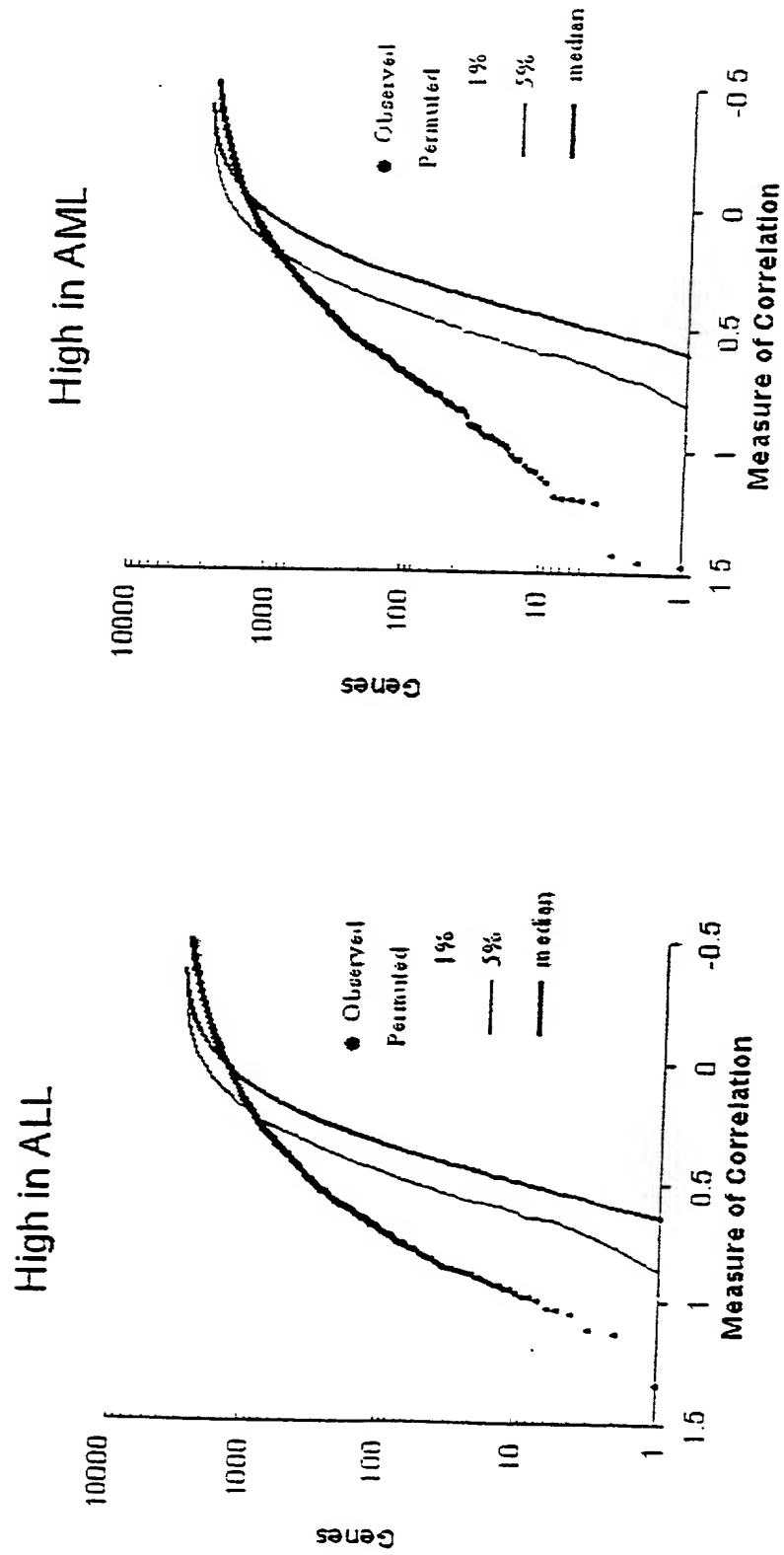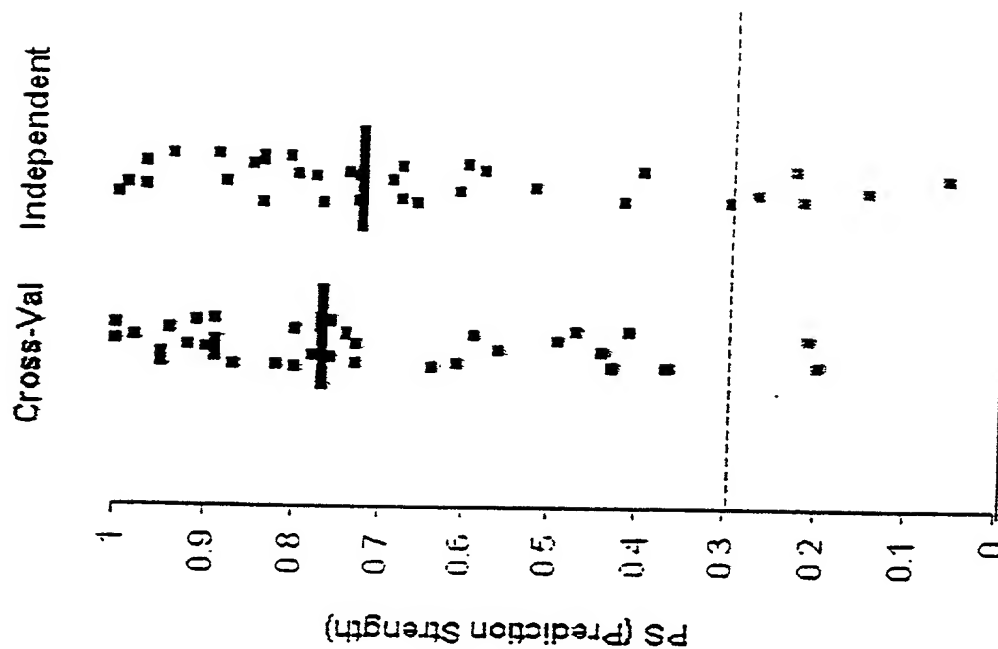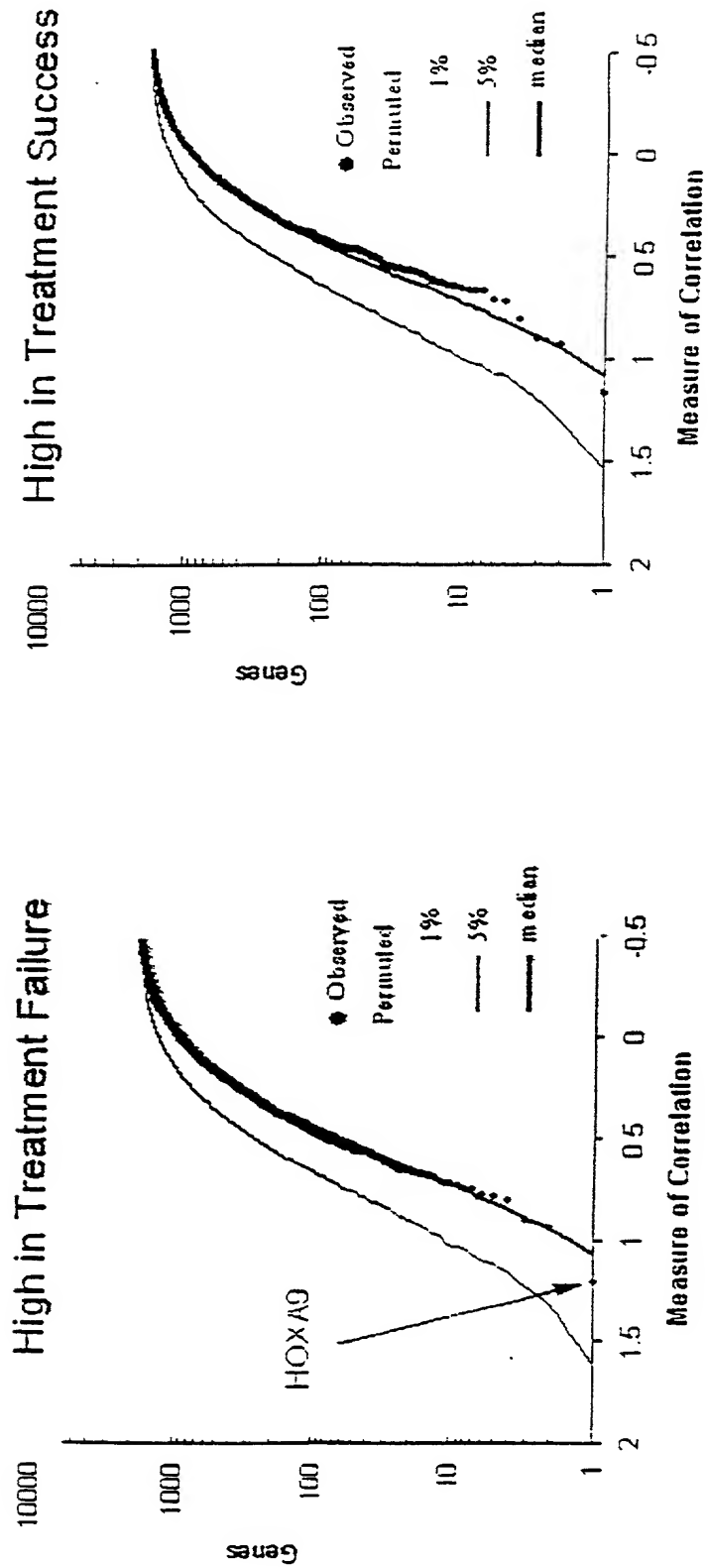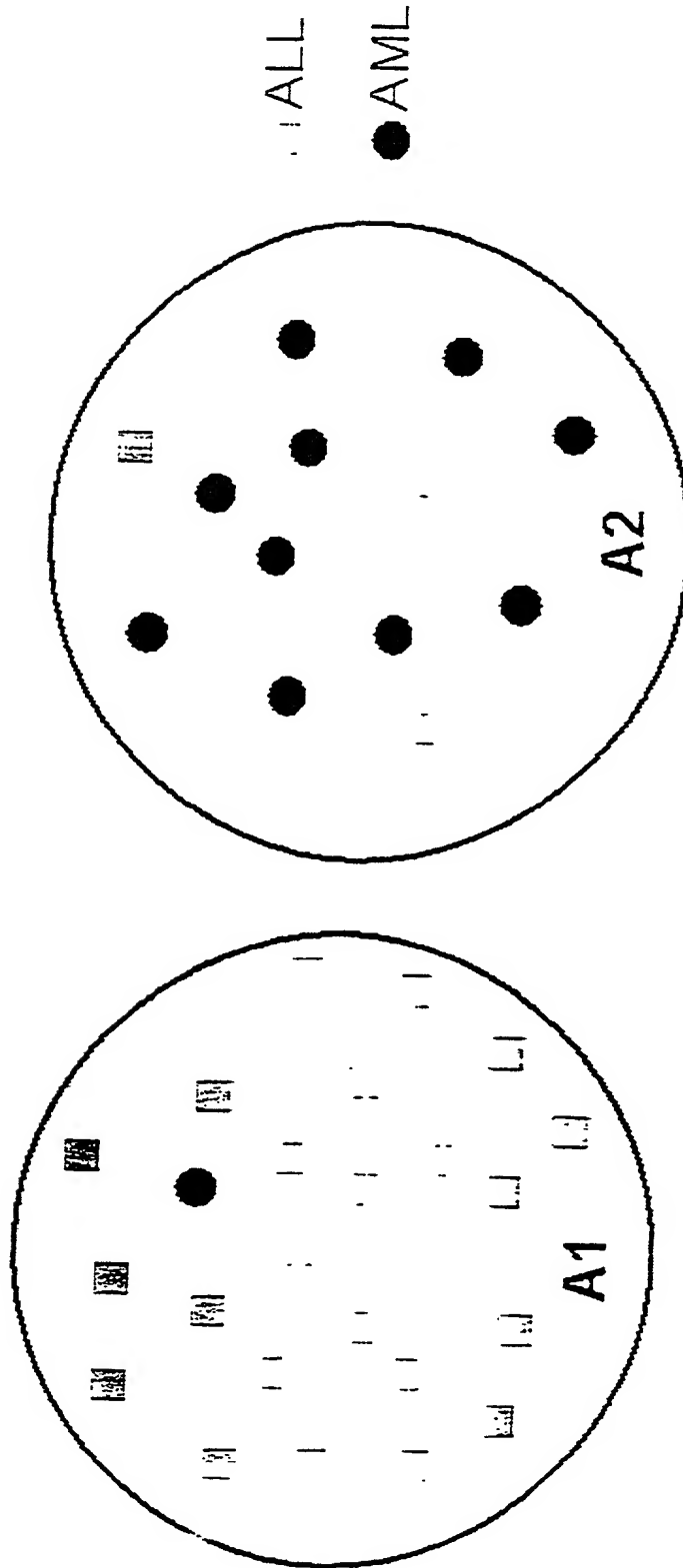|	Weight	ALL	AML
	w_1	v_1	
	w_2	v_2	
	w_3		v_3
	w_4	v_4	
	w_5	v_5	
		. . .	. . .
		$\overline{V_{ALL}}$	$\overline{V_{AML}}$

Figure 1C

Figure 2

Fig. 3A

FIG. 3B

Figure 4

ALL

AML

A2

A1
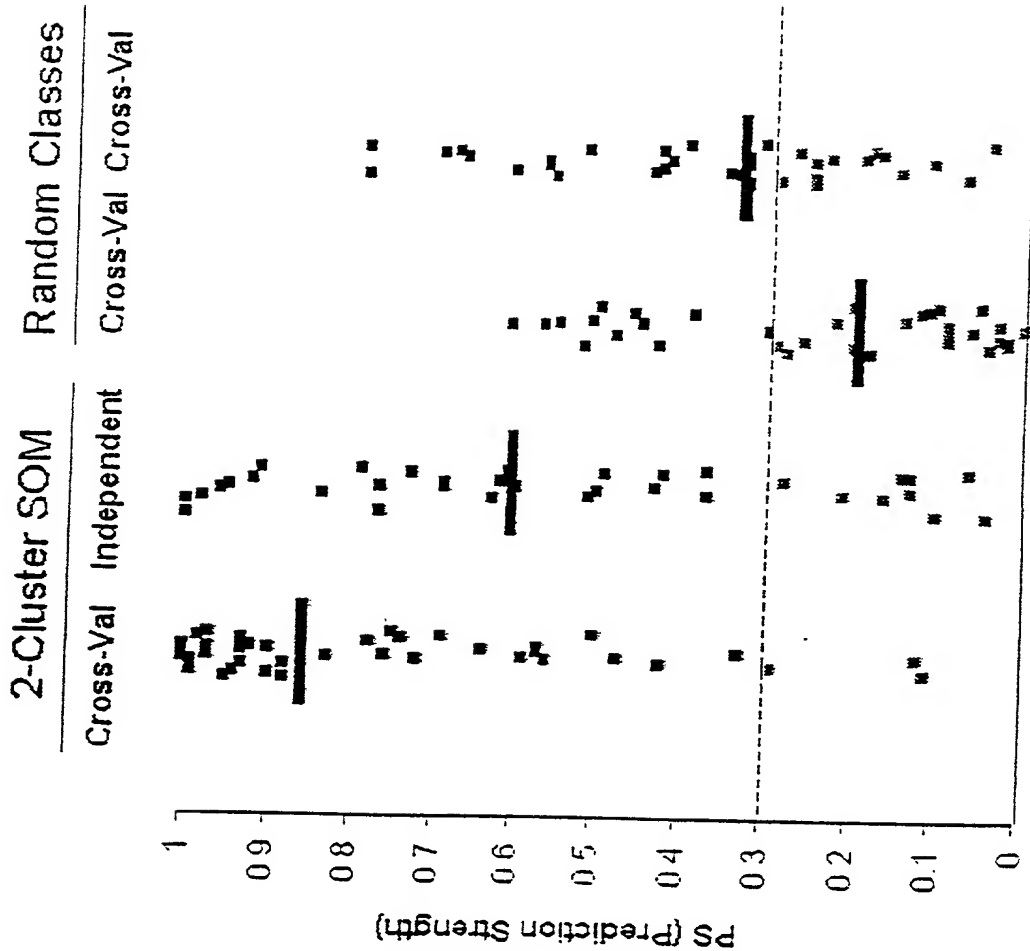
Figure 5A

Figure 5B

Figure 5C

Figure 5D

Input Source 17

I/O Process 19

11

13

CPU 15

Storage 21

I/O Process 23

Output
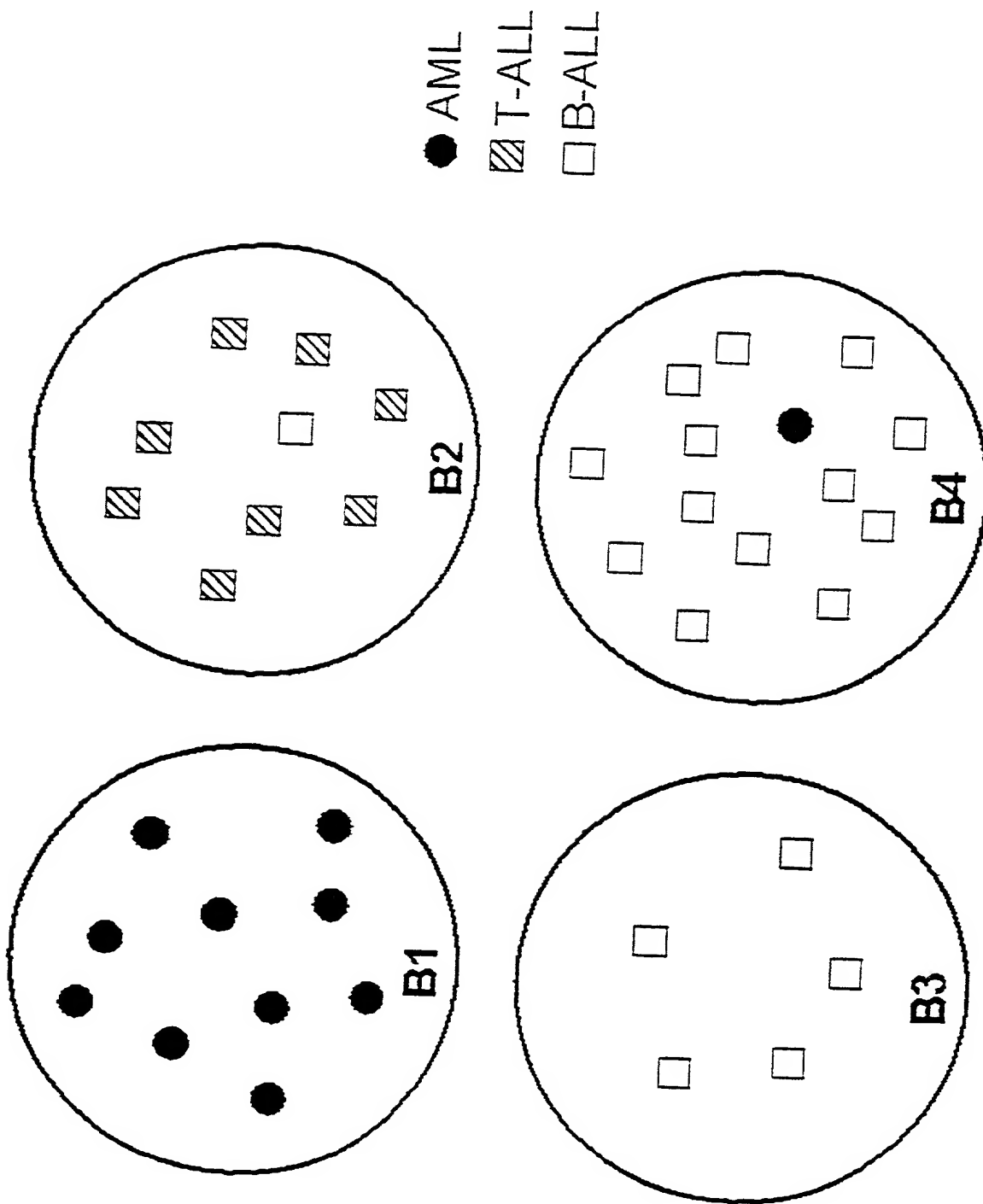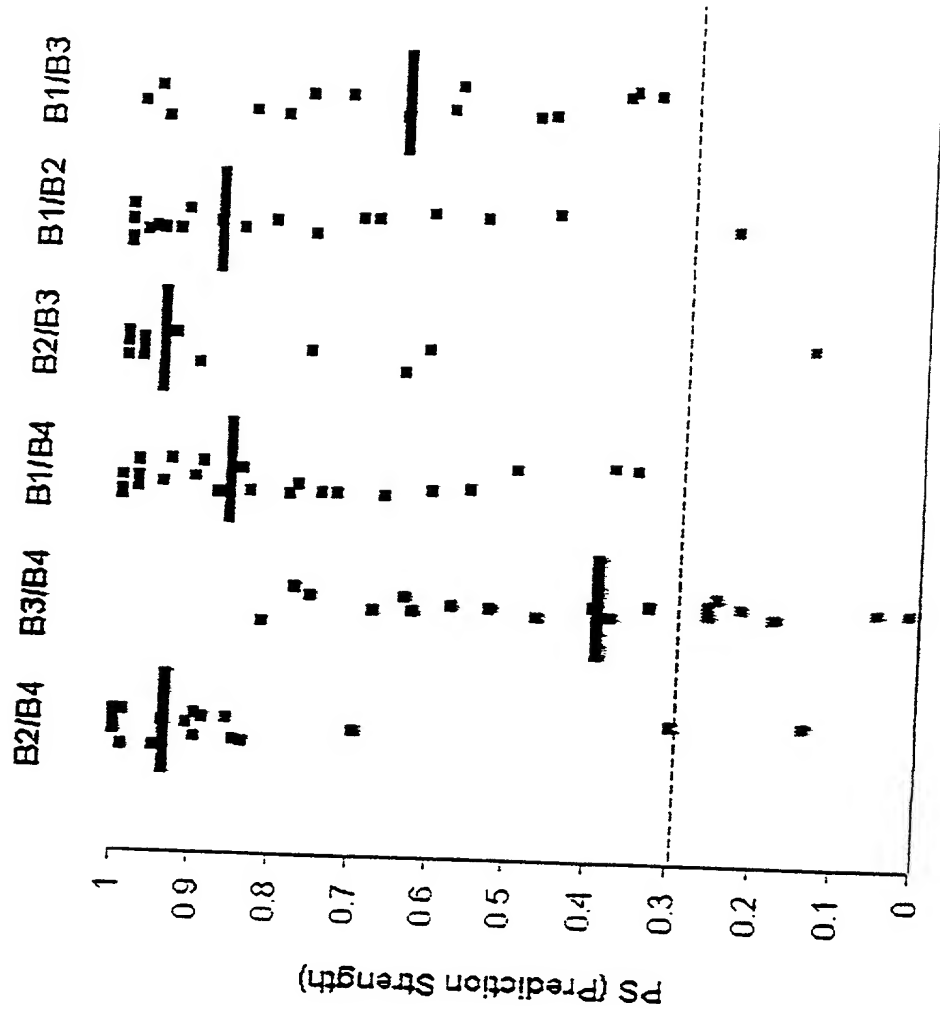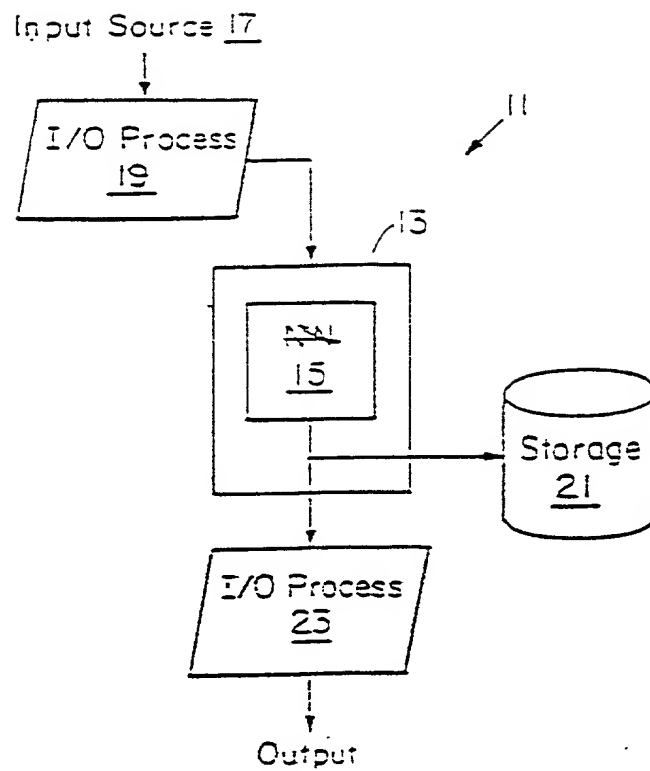
Figure 6

Example of Self-Organizing Map Class Discovery:

Lymphoma Large B-Cell and Follicular

Sample collection:
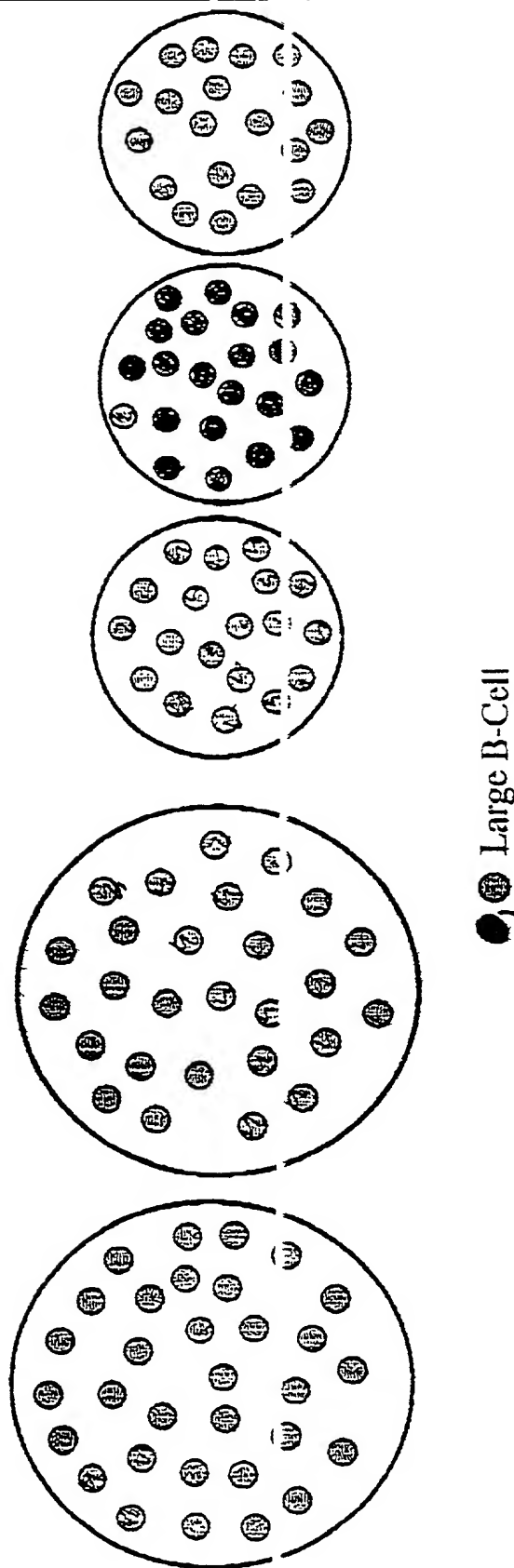
39 large B-Cell + 20 Follicular

Discovered Classes:

●,◉ Large B-Cell

◉ Follicular

Fig. 7

Example of Self-Organizing Map Class Discovery:

Brain Glioma and Medulloblastoma

Sample collection:
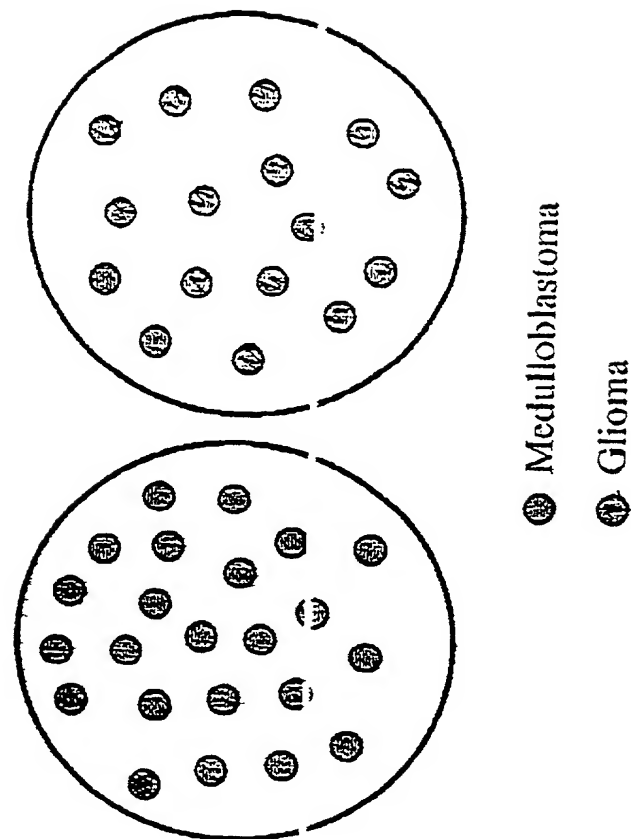
24 Medulloblastomas + 15 Gliomas

Discovered Classes:

● Medulloblastoma

◉ Glioma

Fig. 8

Multidimensional Scaling of Leukemia Samples

(431 genes, 15-fold, delta > 1500, thres=100, ceil=16,000)
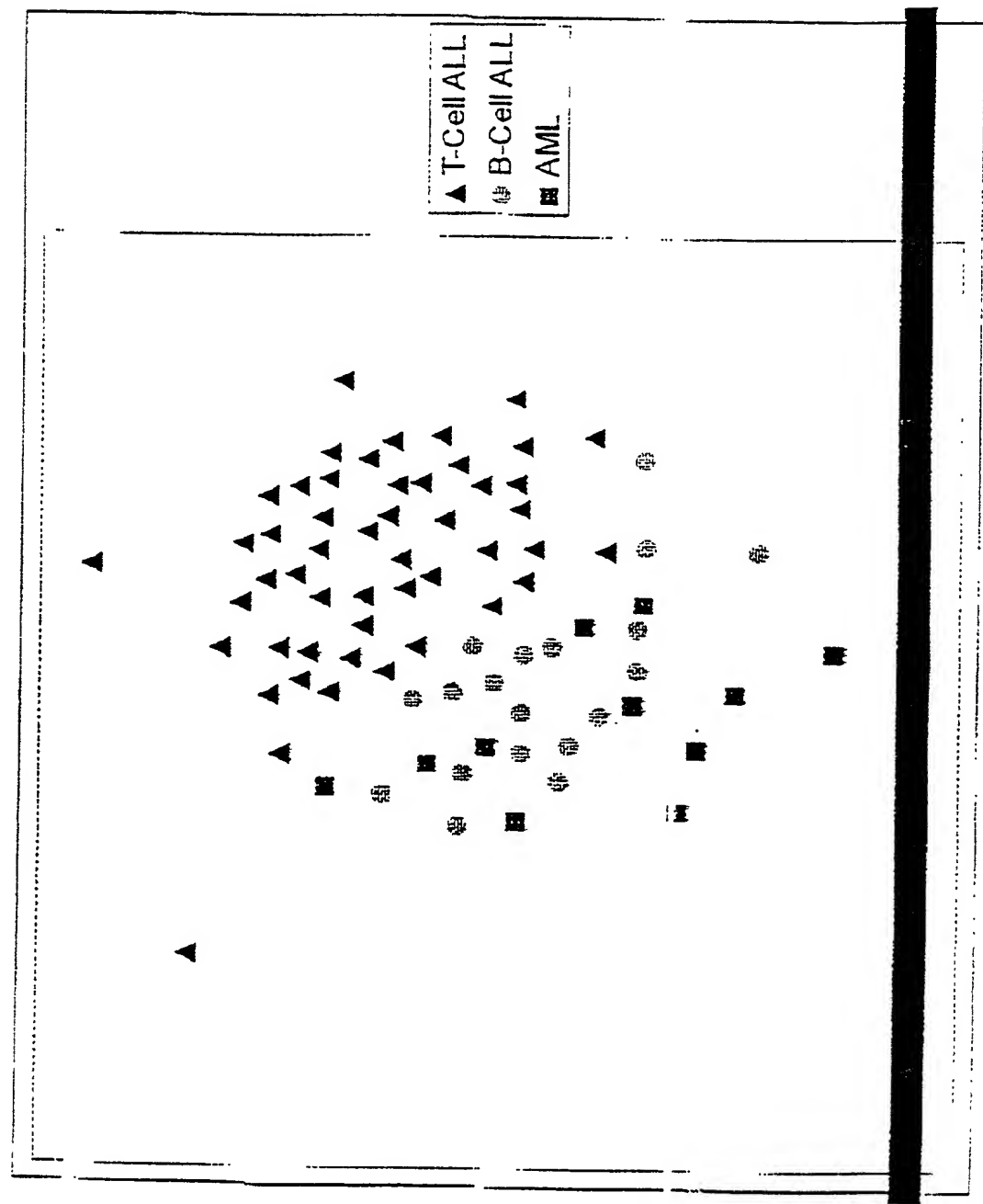

Legend:
▲ T-Cell ALL
⊕ B-Cell ALL
▓ AML

Fig. 9

Hierarchy of Problems in Molecular Class Prediction

Problem:	Difficulty:	Gene Markers:	Error:	Example:
I. Tissue or Cell Type Normal *vs.* Abnormal	Low	~1000-2000	~0%	Normal *vs.* Renal Carcinoma
II. Morphological Type	Low-medium	~200-500	~0-5%	Leukemia ALL *vs.* AML
III. Morphological Subtype	Medium-high	~50-100	~0-15%	ALL B- *vs.* T-Cell
IV. Treatment Outcome Drug Sensitivity	High	~1-20	~5-50%	AML Treatment Outcome

Fig. 10

Neighborhood Analysis: Assessing Statistical Significance of Gene-Class Correlations
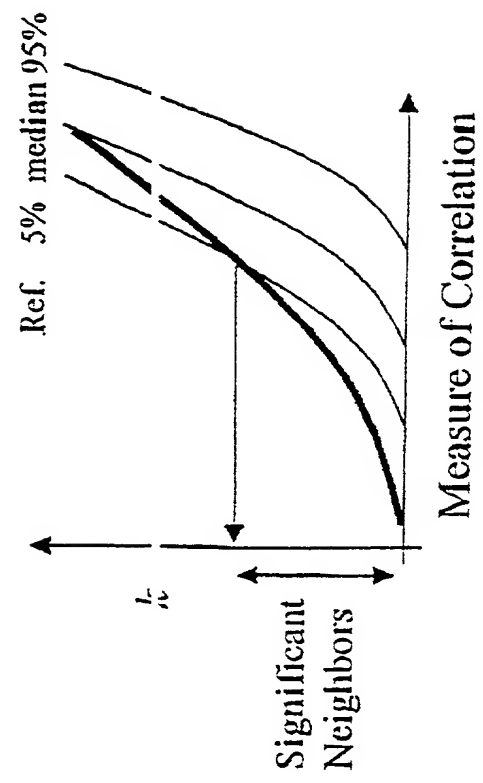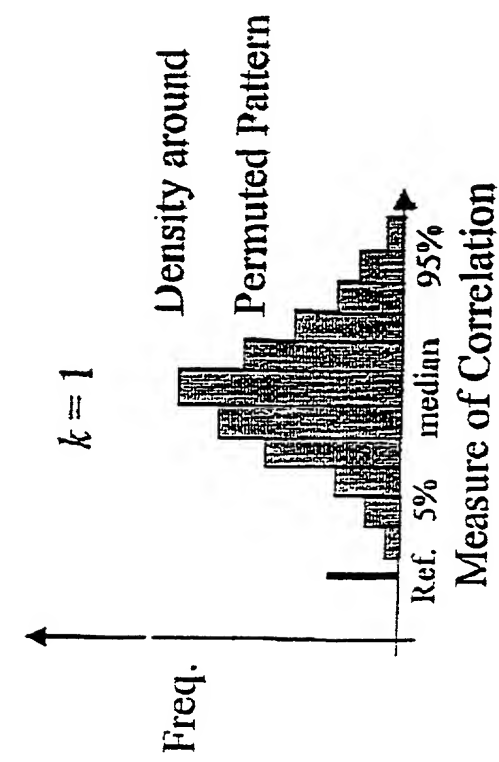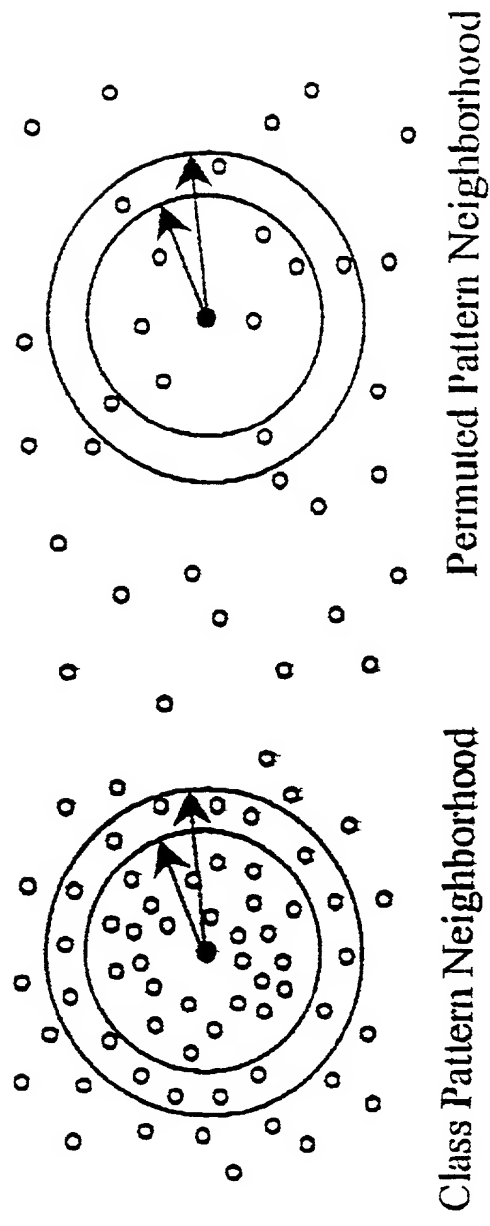
Class Pattern Neighborhood

Permuted Pattern Neighborhood

$k = 1$

Density around
Permuted Pattern

Freq.

Ref.   5%   median   95%
Measure of Correlation

Ref.   5%   median   95%

Significant
Neighbors

Measure of Correlation

Fig. 11

Class Prediction Results

Problem Type	Biological System	Problem Distinction	Number of samples	Number of errors	Number of no calls	Number of errors (all calls)	Number of parameters >100
	Part 1	NormH vs Carcinoma	12	0(0%)	0	0(0%)	>100
I	Leukemia	ALL vs AML	35	0(0%)	2	0(0%)	700
II	Leukemia	ALL B vs T-Cell	33	0(0%)	1	1(3%)	200
NI	Leukemia	Treatment Outcome	15	2(13%)	0	2(13%)	

Fig. 12